CAPACIT - ESTAT

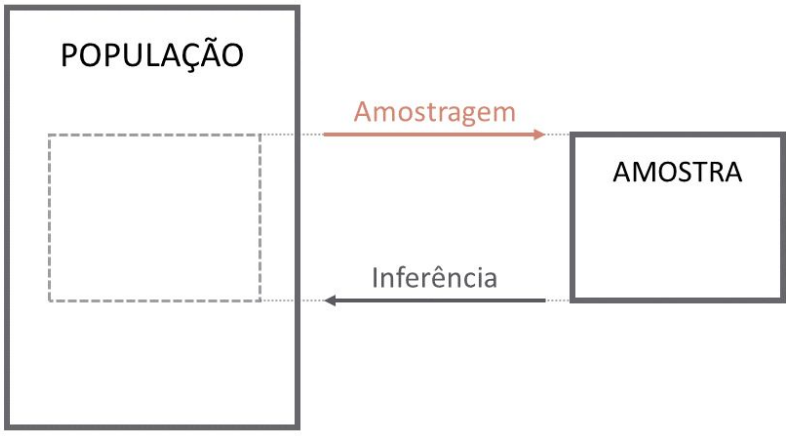
Guilhermo Pastore

2023-07-27

CAPACITAÇÃO - ESTATÍSTICA

Case do concreto

Suponha que somos um empresa interessada em produzir concreto, as matérias-prima fundamentais são cimento, areia, e água. Sabendo disso, a fim de manter a aualidade da produção, é ideal que a quantidade de cimento no processo produtivo esteja em torno de 300 kg/m^3. Históricamente, a média da quantidade de cimento é 300 kg/m^3 e o desvio-padrão 8 kg/m^3. Dado o alto custo e trabalho de coletar informações sobre toda produção, fomos contratados para garantir que a maior do concreto esteja dentro do padrão de qualidade.



Como não temos à toda produção, trabalheremos com uma amostra de 10 cimentos, e a partir faremos uma inferência para toda produção e verificar com certo nível de confiança se o padrão de qualidade está sendo cumprido.

O questionamento seria, como não será possível medir todos os cimentos, como garantir que justamente os que não foram medidos também estão dentro do padrão de qualidade? O principal seria refletir se há propridades o suficiente que torne essa amostra "boa". Quais seriam essas propriedades?

Critérios para teste de hipótese

Ainda nos baseando no case, o critério que usaremos para decidir se a qualidade está sendo cumprida, é se a média populacional é igual a 300 ou diferente de 300. Em estatística, chamamos a hipótese que contém a igualdade de hipótese nula, e a outra de hipótese alternativa. Com isso temos a seguinte situação:

$$\begin{cases} H_0 : \mu = 300 \\ H_a : \mu \neq 300 \end{cases}$$

Porém, como só temos acesso a uma amostra de tamanho 10 (ou seja, não sabemos μ), como será possível realizar essa inferência?

Como o nosso parâmetro de interesse é a média populacional, podemos fazer uma inferência utilizando a média amostral. O que a média amostral pode nos dizer sobre a populacional? No mínimo, medimos informações de parte da população, mas isso não necesssariamente garante que o que não medimos na amostra irá seguir o mesmo padrão.

O TLC (Teorema do Limite Central) nos ajuda a responder essas quesões, dado que, com ele iremos perceber que no geral, a média amostral converge para a populacional, e o jeito mais fácil de observar isso é entendendo a distribuição da média amostral.

Simulação do TLC

Por enquanto, deixem o case do concreto descansando na mente de vocês, agora vamos simular o teorema citado utilizando a base de dados ***gapminder.***

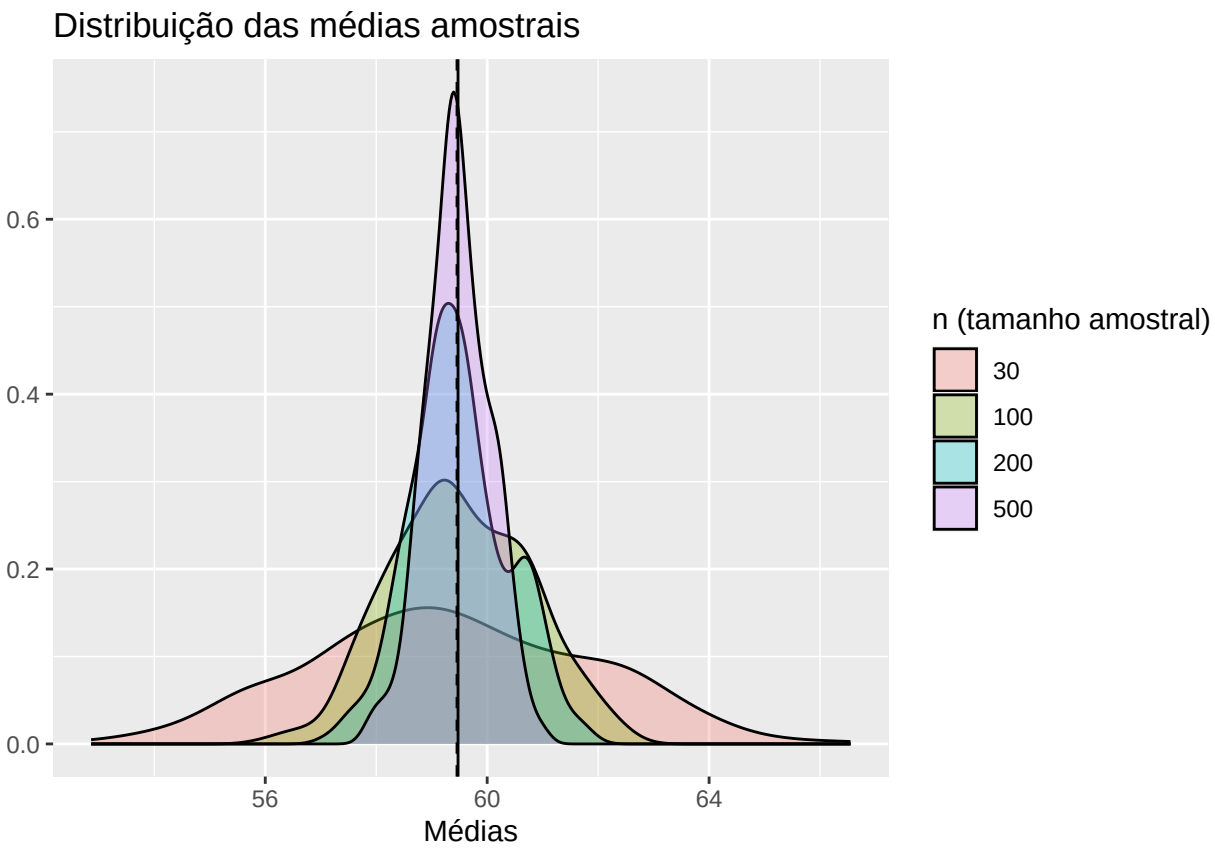
Escolhemos a variável *lifeExp* para essa simulação. Temos acesso à população dessa variável, mas vamos fingir que não, dessa forma, a partir dessa população vamos construir diversas subamostras e ver se o comportamento da média amostral se assemelha ao da populacional.

```
library(tidyverse)
library(gapminder)

df <- gapminder %>%
  mutate(id = row_number(),
         amostra = 0,
         samp_size = 0)

amostras <- df


for (nsample in c(30, 100, 200, 500)){
  for (namostra in 1:300){
    amostras <- amostras %>%
      rbind(df %>%
              mutate(amostra = namostra,
                     samp_size = nsample) %>%
              filter(id %in% sample(1:1704, nsample, replace = TRUE)))
  }
  print(nsample)
}
```

```
## [1] 30
```

```
## [1] 100
## [1] 200
## [1] 500
```

```
amostras %>%
  filter(amostra != 0) %>%
  group_by(amostra, samp_size) %>%
  summarise(media = mean(lifeExp)) %>%
  ggplot(aes(x = media, fill = factor(samp_size))) +
    geom_density(alpha = .3) +
    geom_vline(aes(xintercept = mean(media)), linetype = "dashed") +
    geom_vline(xintercept = mean(df$lifeExp))+
    labs(title = 'Distribuição das médias amostrais', x = 'Médias',
         y = NULL, fill = 'n (tamanho amostral)')
```



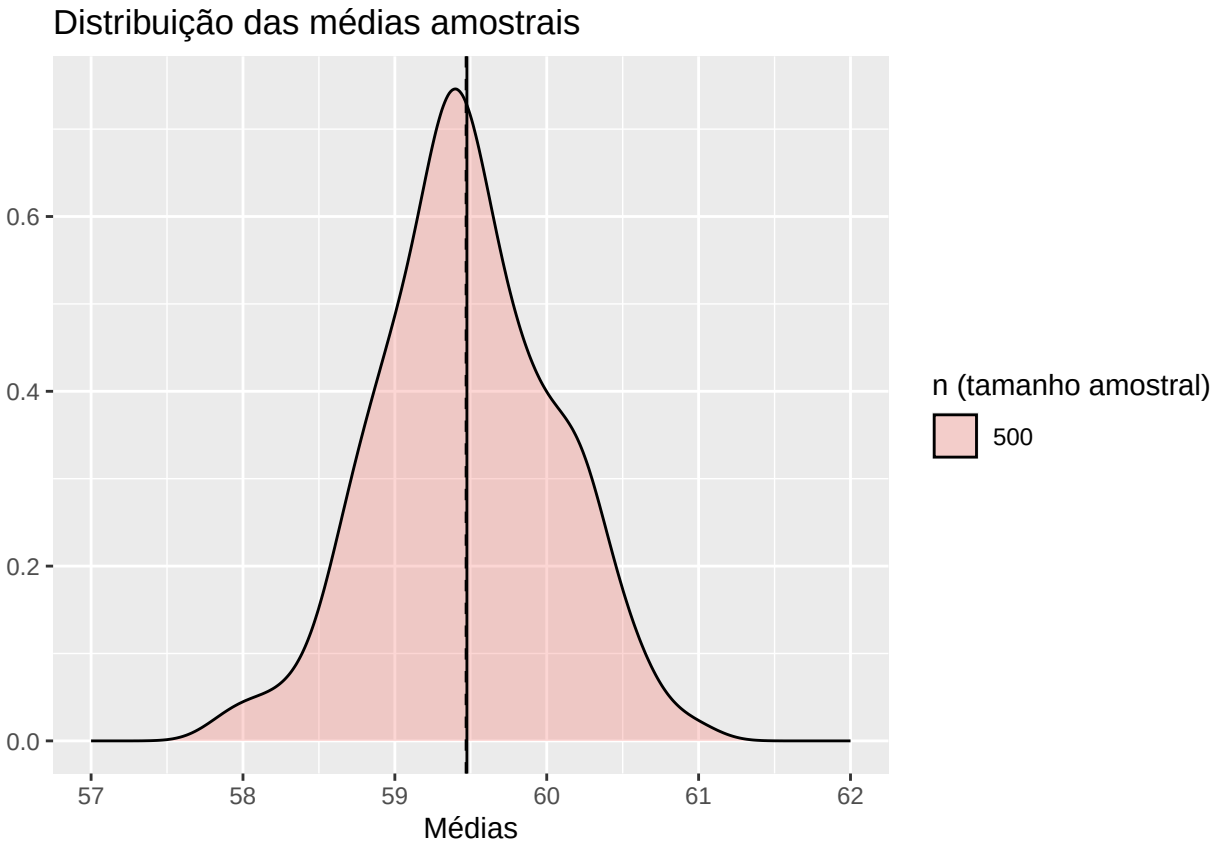
Realizando o teste de hipótese (gapminder)

Vamos supor que queremos verificar se a média da expectativa de vida no mundo é ou não de 60 anos, ou seja:

$$\begin{cases} H_0 : \mu = 60 \\ H_a : \mu \neq 60 \end{cases}$$

```r
amostras %>%
  filter(amostra != 0, samp_size == 500) %>%
  group_by(amostra, samp_size) %>%
  summarise(media = mean(lifeExp)) %>%
  ggplot(aes(x = media, fill = factor(samp_size))) +
    geom_density(alpha = .3) +
    geom_vline(aes(xintercept = mean(media)), linetype = "dashed") +
    geom_vline(xintercept = mean(df$lifeExp))+
    labs(title = 'Distribuição das médias amostrais', x = 'Médias',
         y = NULL, fill = 'n (tamanho amostral)')+
    xlim(57,62)
```

```
## `summarise()` has grouped output by 'amostra'. You can override using the
## `.groups` argument.
```



Vamos supor que, coletamos uma amostra aleatória de tamnho 500, e nela observamos $\bar{x} = 58,2$, além disso, vamos supor que por algum motivo sabemos que o desvio-padrão populacional é igual a 12,91 ($\sigma = 12,91$). Como temos uma distribuição normal, ao invés de trabalhar com a própria média amostral, podemos utilizar a estatística padronizada z. Podemos padronizar a estatística da seguinte forma:

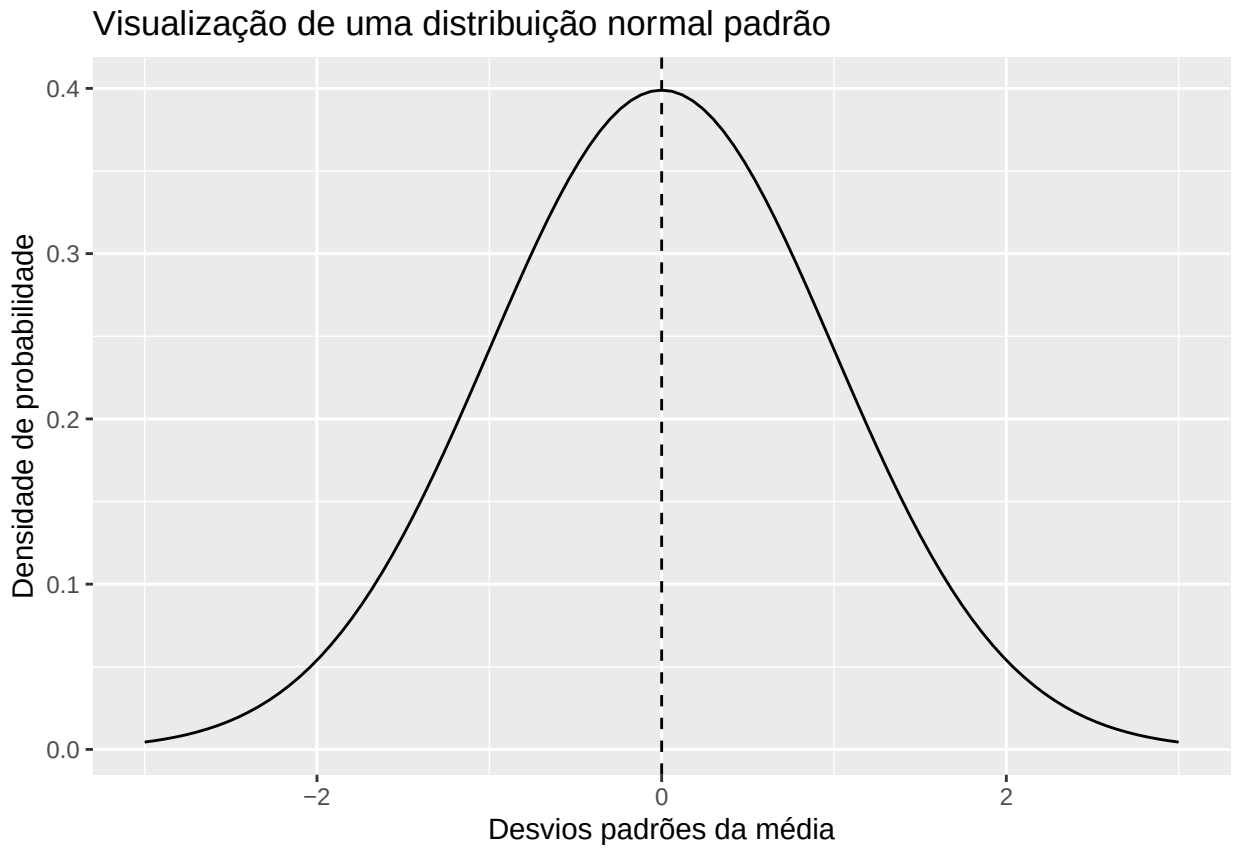
$$z_{obs} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Ou seja, do exemplo temos que:

$$z_{obs} = \frac{58.2 - 60}{\frac{12.9171}{\sqrt{500}}} \approx -3,1176$$

Agora que já temos nossa estatística do teste padronizada, podemos calcular a probabilidade de observar novamente algo tão ou mais desfavorável à hipótese nula. Para isso, vamos utilizar o comando *pnorm*.

```
ggplot(data = data.frame(x = c(-3, 3)), aes(x)) +
  stat_function(fun = dnorm) +
  geom_vline(xintercept = 0, colour = "black", linetype = "dashed") +
  labs(
    title = "Visualização de uma distribuição normal padrão",
    y = "Densidade de probabilidade",
    x = "Desvios padrões da média"
  )
```




```
#CALCULANO O P-VALOR

p_valor <- 2*pnorm(-3.1176)
cat('O p-valor do teste é', p_valor)
```

```
## O p-valor do teste é 0.001823301
```

Dado que a probabilidade de observamos algo parecido com o que já observamos é extremamente baixa, é mais provável que formulamos a hipótese de maneira equivocada do que estarmos observando um milagre, por isso, rejeitamos a hipótese nula. A regra de decisão para um teste de hipótese é:

$$\begin{cases} p - valor < \alpha \rightarrow \text{Rejeita } H_0 \\ p - valor > \alpha \rightarrow \text{Não rejeita } H_0 \end{cases}$$

Sendo α o nível de significância, grosseiramente, com quanto de chance estou disposto a errar no teste, ou seja, se $\alpha = 5\%$, teremos os reultados do teste com 95% de confiança. Usualmente se usa $\alpha = 5\%$, então nesse caso, rejeitamos H_0 já que $0,18\% < 5\%$.

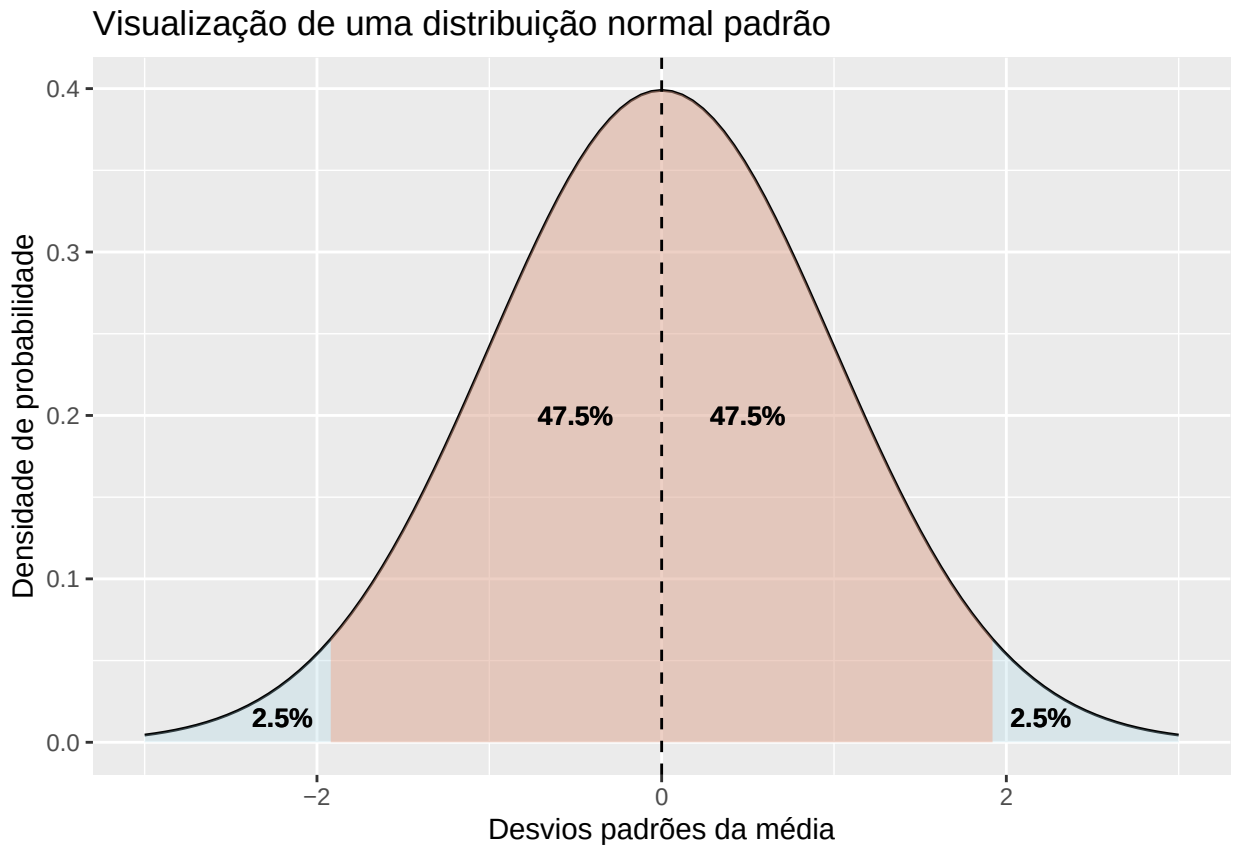
Intervalo de Confiança

Por fim, vamos entender como montar um intervalo de confinça e como interpreta-lo. A fórmula para o intervalo de confiança é:

$$IC[\mu; \gamma] = \left[\bar{x} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right]$$

Para ficar mais claro, vamos fazer uma representação gráfica:

```
library(tidyverse)

dnorm_draw <- function(x, desvios = 1){
  norm_draw <- dnorm(x)
  norm_draw[x <= -desvios | x >= desvios] <- NA
  return(norm_draw)
}

ggplot(data = data.frame(x = c(-3, 3)), aes(x)) +
  stat_function(fun = dnorm) +
  stat_function(
    fun = dnorm,
    geom = "area", fill = "lightblue", alpha = .3) +
  stat_function(
    fun = dnorm_draw, args = list(desvios = qnorm(1-0.025)),
    geom = "area", fill = "coral", alpha = .3
  ) +
  geom_vline(xintercept = 0, colour = "black", linetype = "dashed") +
  geom_text(x = .5, y = .2, size = 3.5, fontface = "bold",
            label = "47.5%") +
  geom_text(x = -.5, y = .2, size = 3.5, fontface = "bold",
            label = "47.5%") +
  geom_text(x = 2.2, y = .015, size = 3.5, fontface = "bold",
            label = "2.5%") +
  geom_text(x = -2.2, y = .015, size = 3.5, fontface = "bold",
            label = "2.5%") +
  labs(
    title = "Visualização de uma distribuição normal padrão",
    y = "Densidade de probabilidade",
    x = "Desvios padrões da média"
  )
```



Podemos observar que, há 2,5% de probabilidade em cada cauda, ou seja, com o intervalor de confiança, estamos interessados que a estimativa esteja nessa região do meio (vermelha) com 95% de probabilidade acumulada. Intuitivamente, se fizermos 100 intervalos a média estarão dentro dos limites delimitados em 95 deles. Por fim, o que queremos dizer é que, se retirarmos outra amostra, há 95% de chance da média dela estar dentro do intervalo delimitado.

De volta ao concreto...

Vamos supor que observamos $\bar{x} = 310$, com isso, podemos calcular a estatística padronizada e tomar uma decisão com 95% de confiança.

$$z_{obs} = \frac{310 - 300}{\frac{8}{\sqrt{10}}} \approx 3,9528$$

```
2*pnorm(3.9528, lower.tail = F)
```

```
## [1] 7.7242e-05
```

$$\text{p-valor} \approx 0\% < 5\% \rightarrow \text{rejeita } H_0$$

$$IC[\mu; \gamma = 95\%] = \left[310 \pm 1,96 * \frac{8}{\sqrt{10}}\right] = [305,0415 \ ; \ 314,9585]$$

E se $\bar{x} = 301$?

$$z_{obs} = \frac{301 - 300}{\frac{8}{\sqrt{10}}} \approx 0,3952$$

```
2*pnorm(0.3952, lower.tail = F)
```

```
## [1] 0.6926953
```

$$\text{p-valor} \approx 69,26\% > 5\% \rightarrow \text{não rejeita } H_0$$

$$IC[\mu; \gamma = 95\%] = \left[301 \pm 1,96 * \frac{8}{\sqrt{10}}\right] = [296,0415 \ ; \ 305,9585]$$

Último exemplo...

Utilizando a base *mtcars*, suponha que estamos interessado em descobrir se o tipo de tranmissão de um carro influencia o quão econômico ele é. Para testar isso, podemos verificar se a média das milhas por galão dos carros manuais e automáticos são iguais ou diferentes, ou seja:

$$\begin{cases} H_0 : \mu_m = \mu_a \\ H_a : \mu_m \neq \mu_a \end{cases}$$

Antes de realizar esse teste, precisamos verificar se as variâncias são as mesmas entre os dois grupos:

$$\begin{cases} H_0 : \sigma_m^2 = \sigma_a^2 \\ H_a : \sigma_m^2 \neq \sigma_a^2 \end{cases}$$

```
var.test(mpg~am, data = mtcars)
```

```
##
##  F test to compare two variances
##
## data:  mpg by am
## F = 0.38656, num df = 18, denom df = 12, p-value = 0.06691
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1243721 1.0703429
## sample estimates:
## ratio of variances
##          0.3865615
```

Dados que o p-valor é maior do que 5%, temos evidências para não rejeitar H_0, ou seja, as variâncias são iguais entre os dois grupos. Sabendo disso, vamos realizar o teste para as médias.

```
t.test(mpg~am, data = mtcars, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  mpg by am
## t = -4.1061, df = 30, p-value = 0.000285
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -10.84837  -3.64151
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Dado que o p-valor é quase 0%, temos evidências para rejeitar H_0, ou seja, as médias são diferentes entre os tipos de carro.